On Pairs of f-divergences and their Joint Range

Peter Harremoës, *Member, IEEE,* Igor Vajda†, *Fellow IEEE*

***Abstract*—We compare two f-divergences and prove that their joint range is the convex hull of the joint range for distributions supported on only two points. Some applications of this result are given.**

***Index Terms*—f-divergence, convexity, joint range.**

I. Divergences and divergence statistics

MANY of the divergence measures used in statistics are of the f-divergence type introduced independently by I. Csiszár [1], T. Morimoto [2], and Ali and Silvey [3]. Such divergence measures have been studied in great detail in [4]. Often one is interested inequalities for one f-divergence in terms of another f-divergence. Such inequalities are for instance needed in order to calculate the relative efficiency of two f-divergences when used for testing goodness of fit but there are many other applications. In this paper we shall study the more general problem of determining the joint range of any pair of f-divergences. The results are useful in determining general conditions under which information divergence is a more efficient statistic for testing goodness of fit than another f-divergence, but will not be discussed in this short paper.

Let $f : (0, \infty) \to \mathbb{R}$ denote a convex function satisfying $f(1) = 0$. We define $f(0)$ as the limit $\lim_{t\to 0} f(t)$. We define $f^*(t) = tf\left(t^{-1}\right)$. Then f^* is a convex function and $f^*(0)$ is defined as $\lim_{t\to 0} tf\left(t^{-1}\right) = \lim_{t\to\infty} \frac{f(t)}{t}$. Assume that P and Q are absolutely continuous with respect to a measure μ, and that $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$. For arbitrary distributions P and Q the f-divergence $D_f(P,Q) \geq 0$ is defined by the formula

$$D_f(P,Q) = \int_{\{q>0\}} f\left(\frac{p}{q}\right)\, dQ + f^*(0)\, P(q=0) \qquad (1)$$

(for details about the definition (1) and properties of the f-divergences, see [5], [4] or [6]). With this definition

$$D_f(P,Q) = D_{f^*}(Q,P).$$

Example 1: The function $f(t) = |t-1|$ defines the L^1-distance

$$\|P - Q\| = \sum_{j=1}^{k} q_j \left|\frac{p_j}{q_j} - 1\right| = \sum_{j=1}^{k} |p_j - q_j| \quad \text{(cf. (1))} \qquad (2)$$

which plays an important role in information theory and mathematical statistics [7], [8] .

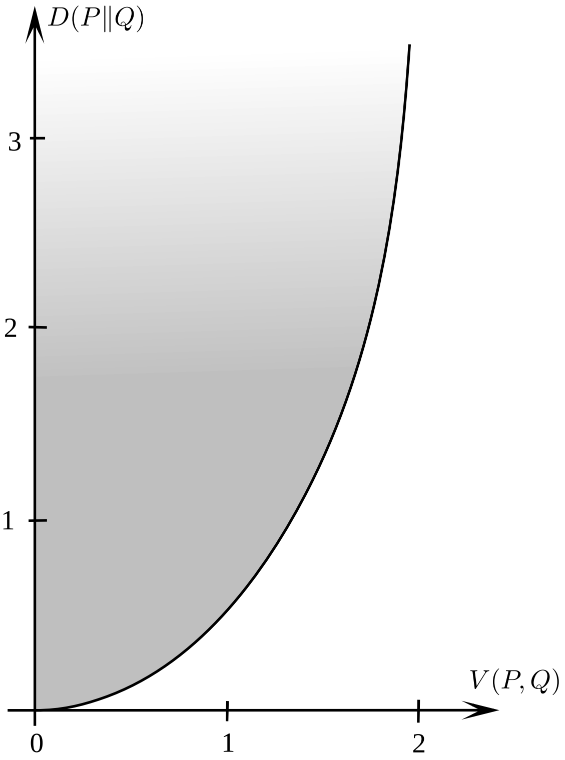


Fig. 1. The joint range of total variation V and information D as determined in [8]. It was also proved that any point in the range

In (1) is often taken the convex function f which is one of the power functions ϕ_α of order $\alpha \in \mathbb{R}$ given in the domain $t > 0$ by the formula

$$\phi_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)} \quad \text{when} \quad \alpha(\alpha-1) \neq 0 \qquad (3)$$

and by the corresponding limits

$$\phi_0(t) = -\ln t + t - 1 \quad \text{and} \quad \phi_1(t) = t\ln t - t + 1. \qquad (4)$$

The ϕ-divergences

$$D_\alpha(P,Q) \overset{def}{=} D_{\phi_\alpha}(P,Q), \quad \alpha \in \mathbb{R} \qquad (5)$$

based on (3) and (4) are usually referred to as power divergences of orders α. For details about the properties of power divergences, see [5] or [6]. Next we mention the best known members of the family of statistics (5), with a reference to the skew symmetry $D_\alpha(P,Q) = D_{1-\alpha}(Q,P)$ of the power divergences (5).

Manuscript received xxxxx, 2010; revised xxxxx. This work was supported by the European Network of Excellence and the GAČR grants 102/07/1131 and 202/10/0618.

Peter Harremoës is with Copenhagen business College, Denmark. Igor Vajda passed away during the preparation of this article. He worked at Institute of Information and Automation, Prague, Czech Republic.

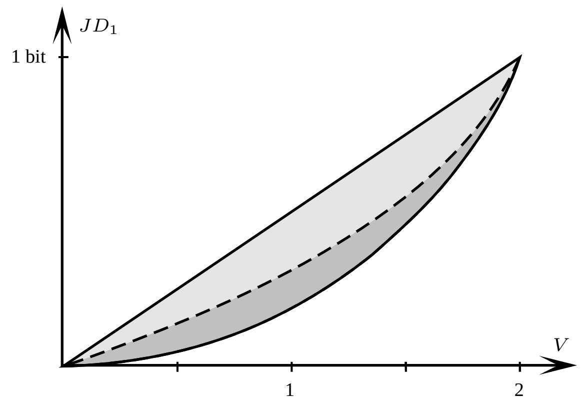


Fig. 2. Joint range of total variation and Jensen-Shannon divergence. The 2-point achievable pairs have dark shading and the 3-point achievable pairs have light shading.

Example 2: The χ^2-divergence (or quadratic divergence or Pearson divergence)

$$D_2(P,Q) = D_{-1}(Q,P) = \frac{1}{2}\sum_{j=1}^{k}\frac{(p_j-q_j)^2}{q_j} \tag{6}$$

leads to the well known Pearson and Neyman statistics. The information divergence

$$D_1(P,Q) = D_0(Q,P) = \sum_{j=1}^{k} p_j \ln\frac{p_j}{q_j} \tag{7}$$

leads to the log-likelihood ratio and reversed log-likelihood ratio statistics. The symmetric Hellinger divergence

$$D_{1/2}(P,Q) = D_{1/2}(Q,P) = H(P,Q)$$

leads to the Freeman–Tukey statistic.

Example 3: The Hellinger divergence and the total variation are symmetric in the arguments P and Q. Non-symmetric divergences may be symmetrized. For instance the LeCam divergence is nothing but the symmetrized χ^2-divergence given by

$$D_{LeCam}(P,Q) = \frac{1}{2}D_2\left(P,\frac{P+Q}{2}\right) + \frac{1}{2}D_2\left(Q,\frac{P+Q}{2}\right)$$

Another symmetrized divergence is the Jensen Shannon divergence defined by

$$JD_1(P,Q) = \frac{1}{2}D\left(P\left\|\frac{P+Q}{2}\right.\right) + \frac{1}{2}D\left(Q\left\|\frac{P+Q}{2}\right.\right).$$

The joint range of total variation with Jensen Shannon divergence was studied by Briët and Harremoës [9] and is illustrated on Figure 2.

In this paper we shall prove that the joint range of any pair of f-divergences is essentially determined by the range of distributions on a two-element set. In special cases the significance of determining the range over two-element set has been pointed out explicitly in [10]. Here we shall prove that a reduction to two-element sets can always be made.

II. Joint range of f-divergences

In this section we are interested in the range of the map $(P,Q) \to (D_f(P,Q), D_g(P,Q))$ where P and Q are probability distributions on the same set.

Definition 4: A point $(x,y) \in \mathbb{R}^2$ is (f,g)-*achievable* if there exist probability measures P and Q on a σ-algebra such $(x,y) = (D_f(P,Q), D_g(P,Q))$. A (f,g)-divergence pair (x,y) is d-*achievable* if there exist probability vectors $P,Q \in \mathbb{R}^d$ such that

$$(x,y) = (D_f(P,Q), D_g(P,Q)).$$

Lemma 5: Assume that

$$P_0(A) = Q_0(A) = 1$$

and

$$P_1(B) = Q_1(B) = 1$$

and that $A \cap B = \varnothing$. If $P_\alpha = (1-\alpha)P_0 + \alpha P_1$ and $Q_\alpha = (1-\alpha)Q_0 + \alpha Q_1$ then

$$D_f(P_\alpha,Q_\alpha) = (1-\alpha)D_f(P_0,Q_0) + \alpha D_f(P_1,Q_1).$$

Theorem 6: The set of (f,g)-achievable points is convex.

Proof: Assume that (P,Q) and $\left(\tilde{P},\tilde{Q}\right)$ are two pairs of probability distributions on a space $(\mathcal{X},\mathcal{F})$. Introduce a two-element set $B = \{0,1\}$ and the product space $\mathcal{X}\times B$ as a measurable space. Let ϕ denote projection on B. Now we define a pair $\left(\tilde{P},\tilde{Q}\right)$of joint distribution on $\mathcal{X}\times B$. The marginal distribution of both $\tilde{P}$ is $\tilde{Q}$ on B is $(1-\alpha,\alpha)$. The conditional distributions are given by $P(\cdot \mid \phi = i) = P_i$ and $Q(\cdot \mid \phi = i) = Q_i$ where $i = 0,1$. Then

$$\begin{aligned}\begin{pmatrix} D_f(P_\alpha,Q_\alpha) \\ D_g(P_\alpha,Q_\alpha)\end{pmatrix} &= \\ &\begin{pmatrix}(1-\alpha)D_f(P_0,Q_0)+\alpha D_f(P_1,Q_1) \\ (1-\alpha)D_g(P_0,Q_0)+\alpha D_g(P_1,Q_1)\end{pmatrix} \\ &= (1-\alpha)\begin{pmatrix} D_f(P_0,Q_0) \\ D_g(P_0,Q_0)\end{pmatrix} + \alpha\begin{pmatrix} D_f(P_1,Q_1) \\ D_g(P_1,Q_1)\end{pmatrix} \\ &= (1-\alpha)\begin{pmatrix} D_f(P,Q) \\ D_g(P,Q)\end{pmatrix} + \alpha\begin{pmatrix} D_f\left(\tilde{P},\tilde{Q}\right) \\ D_g\left(\tilde{P},\tilde{Q}\right)\end{pmatrix}.\end{aligned}$$

■

Example 7: For the joint range of total variation and Jensen Shannon divergence illustrated on Figure 2 the set of 2-achievable points is not convex but the set of 3-achievable points is convex and equals the set of all (f,g)-achievable points.

Theorem 8: Any (f,g)-achievable points is a convex combination of two 2-achievable points. Consequently, any (f,g)-achievable point is 4-achievable.

Proof: Let P and Q denote probability measures on Borel space. Define the set $A = \{q > 0\}$ and the function $X = p/q$ on A. Then Q satisfies

$$Q(A) = 1, \tag{8}$$

$$\int_A X \; dQ \leq 1.$$

Now we fix X and A. The formulas for the divergences become

$$\begin{aligned} D_f(P,Q) &= \int_A f(X)\ dQ + f^*(0)\,P\left(\complement A\right) \\ &= \int_A f(X)\ dQ + f^*(0)\left(1-\int_A X\ dQ\right) \\ &= \int_A \left(f(X)\ + f^*(0)(1-X)\right)\ dQ \\ &= \mathrm{E}\left[f(X)+f^*(0)(1-X)\right] \end{aligned}$$

and similarly

$$D_g(P,Q) = \mathrm{E}\left[g(X)\ \ + g^*(0)(1-X)\right].$$

Hence, the divergences only depend on the distribution of X. Therefore we may without loss of generality assume that Q is a probability measure on $[0,\infty)$.

Define C as the set of probability measures on $[0,\infty)$ satisfying $\mathrm{E}[X]\leq 1$. Let C^+ be the set of additive measures μ on $[0,\infty)$ satisfying $\mu(A)\leq 1$ and $\int_A X\ d\mu \leq 1$. Then C^+ is convex and thus compact under setwise convergence. According to the Choquet–Bishop–de Leeuw theorem [11, Sec. 4] any other point in C^+ is the barycenter of a probability measure over such extreme points. In particular an element $Q\in C$ is the barycenter of a probability measure P_{bary} over extreme points of C^+ and these extreme points must in addition be probability measures with P_{bary}-probability 1. Hence $Q\in C$ is a barycenter of a probability measure over extreme points in C.

Let Q be an element in C. Let $A_i, i=1,2,3$ be a disjoint cover of $[0,\infty)$ and assume that $Q(A_i)>0$. Then

$$Q=\sum_{i=1}^{3} Q(A_i)\,Q(\cdot\mid A_i)\,.$$

For a probability vector $\lambda=(\lambda_1,\lambda_2,\lambda_2)$ let Q_λ denote the distribution

$$Q_\lambda=\sum_{i=1}^{3}\lambda_i Q(\cdot\mid A_i)\,.$$

Then Q_λ is element in C if and only if

$$\sum_{i=1}^{3}\lambda_i\int_A X\ dQ(\cdot\mid A_i)\leq 1. \tag{9}$$

An extreme probability vector λ that satisfies (9) has one or two of its weights equal to 0. Hence, if Q is extreme in C and $A_i, i=1,2,3$ is a disjoint cover of A, then at least one of the three sets satisfies $Q(A_i)=0$. Therefore an extreme point $Q\in C$ is of one of the following two types:

1) Q is concentrated in one point.
2) Q has support on two points. In this case the inequality $\int_A X\ dQ\leq 1$ holds with equality and $P(A)=1$ so that P is absolutely continuous with respect to Q and therefore supported by the same two-element set.

The formulas for divergence are linear in Q. Hence any (f,g)-divergence pair is a the barycenter of a probability measure P_{bary} over points generated by extreme distributions $Q\in C$. The extreme distributions of type 2 generate 2-achievable points.

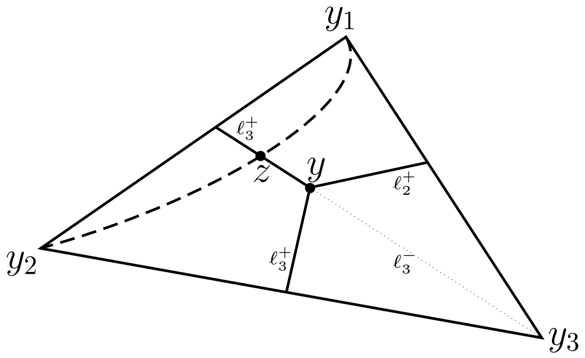
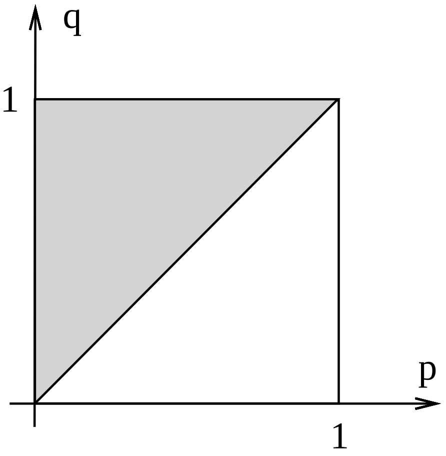


Fig. 3. The slashed curve connects $\mathbf{y}_1$ and $\mathbf{y}_2$. The lines ℓ_1^- and ℓ_2^- are not illustrated.

For extreme points Q concentrated in a single point we can reverse the argument at make a barycentric decomposition with respect to P. If an extreme P has a two-point support then Q is absolutely continuous with respect to P and generates a (f,g)-achievable point that is 2-achievable. If P is concentrated in a point then this point may either be identical with the support of Q and the two probability measures are identical, or the support points are different and P and Q are singular but still (P,Q) is supported on two points. Therefore any (f,g)-achievable point has a barycentric decomposition into 2-achievable points.

Let $\mathbf{y}=(y,z)$ be a (f,g)-achievable point. As we have seen $\mathbf{y}$ is a barycenter of (f,g)-achievable points that are 2-achievable. According to the Carathéodory's theorem [12] any barycentric decomposition in two dimensions may be obtained as a convex combination of at most three points $\mathbf{y}_i,\ i=1,2,3.$ as illustrated in Figure 3. Assume that all three points have positive weight. Let ℓ_i be the line through $\mathbf{y}$ and $\mathbf{y}_i$. The point $\mathbf{y}$ divides the line ℓ_i in two half-lines ℓ_i^+ and ℓ_i^- , where ℓ_i^- denotes the half-line that contains $\mathbf{y}_i$. The lines $\ell_i^+, i=1,2,3$ divide $\mathbb{R}^2$ into three sectors, each of them containing one of the points $\mathbf{y}_i, i=1,2,3$. The set of (f,g)-divergence pairs that are 3-achievable is curve-connected so there exist a continuous curve of (f,g)-divergence pairs that are 2-achievable from $\mathbf{y}_1$ to $\mathbf{y}_2$ that must intersect $\ell_1^+\cup\ell_3^+$ in a point $\mathbf{z}$. If $\mathbf{z}$ lies on ℓ_i^+ then $\mathbf{y}$ is a convex combination of the two points $\mathbf{y}_i$ and $\mathbf{z}$. Hence, any (f,g)-divergence pair is a convex combination of two points that are 2-achievable. From the construction in the proof of Theorem 6 we see that any (f,g)-divergence pair is 4-achievable.

An f-divergence on an arbitrary σ-algebra can be approximated by the f-divergence on its finite sub-algebras. Any finite σ-algebra is a Borel σ-algebra for a discrete space so for probability measures P,Q on a σ-algebra the point $(D_f(P,Q), D_g(P,Q))$ is in the closure of 4-achievable points. For any function pairs (f,g) the intersection of the set of 2-achievable points and the first quadrant is closed. 4-achievable points are convex combinations of 2-achievable points so the intersection of the 4-achievable points and the first quadrant is closed contains $(D_f(P,Q), D_g(P,Q))$ even if P,Q are measures on a non-atomic σ-algebra. ■



The set of (f,g)-achievable points that are 2-achievable can be parametrized as $P=(1-p,p)$ and $Q=(1-q,q)$. If we define $\overline{(1-p,p)}=(p,1-p)$ then $D_f(P,Q)=D_f\left(\overline{P},\overline{Q}\right)$. Hence we may assume without loss of generality assume that $p\leq q$ and just have to determine the image of the simplex $\Delta=\{(p,q)\mid 0\leq p\leq q\leq 1\}$. This result makes it very easy to make a numerical plot of the (f,g)-achievable point is 2-achievable and the joint range is just the convex hull.

III. Image of the triangle

In order to determine the image of the triangle Δ we have to check what happens at inner points and what happens at or near the boundary. Most inner points are mapped into inner points of the range. On subsets of Δ where the derivative matrix is non-singular the mapping $(P,Q)\to(D_f,D_g)$ is open according to the open mapping theorem from calculus. Hence, all inner points that are not mapped into interior points of the range must satisfy

$$\begin{vmatrix} \frac{\partial D_f}{\partial p} & \frac{\partial D_g}{\partial p} \\ \frac{\partial D_f}{\partial q} & \frac{\partial D_g}{\partial q} \end{vmatrix}=0.$$

Depending on functions f and g this equation may be easy or difficult to solve, but in most cases the solutions will lie on a 1-dimensional manifold that will cut the triangle Δ into pieces, such that each piece is mapped isomorphically into subsets of the range of $(P,Q)\to(D_f,D_g)$. Each pair of functions (f,g) will require its own analysis.

The diagonal $p=q$ in Δ is easy to analyze. It is mapped into $(D_f,D_g)=(0,0)$.

Lemma 9: If $f(0)=\infty$, and $\lim_{t\to 0}\inf\frac{g(t)}{f(t)}=\beta_0$, then the supremum of

$$\beta\cdot D_f(P,Q)-D_g(P,Q)$$

over all distributions P,Q is ∞ if $\beta>\beta_0$.

If $f^*(0)=\infty$, and $\lim_{t\to\infty}\inf\frac{g(t)}{f(t)}=\beta_0$, then the supremum of

$$\beta\cdot D_f(P,Q)-D_g(P,Q)$$

over all distributions P,Q is ∞ if $\beta>\beta_0$.

If $g(0)=\infty$, and $\lim_{t\to 0}\sup\frac{g(t)}{f(t)}=\gamma_0$, then the supremum of

$$D_g(P,Q)-\gamma D_f(P,Q)$$

over all distributions P,Q is ∞ if $\gamma<\gamma_0$.

If $g^*(0)=\infty$, and $\lim_{t\to\infty}\sup\frac{g(t)}{f(t)}=\gamma_0$, then the supremum of

$$D_g(Q,P)-\gamma D_f(Q,P)$$

over all distributions P,Q is ∞ if $\gamma<\gamma_0$.

Proof: Assume that

$$f(0)=\infty \text{ and } \lim_{t\to 0}\inf\frac{g(t)}{f(t)}=\beta_0.$$

The first condition implies

$$D_f((1,0),(1/2,1/2))=\infty$$

and the second condition implies that $g(0)=\infty$ and

$$D_g((1,0),(1/2,1/2))=\infty.$$

We have

$$\begin{aligned}\frac{D_g((p,1-p),(1/2,1/2))}{D_f((p,1-p),(1/2,1/2))} &=\frac{g(2p)/2+g(2(1-p))/2}{f(2p)/2+f(2(1-p))/2}\\ &=\frac{g(2p)+g(2(1-p))}{f(2p)+f(2(1-p))}.\end{aligned}$$

Let $(t_n)_n$ be a sequence such that $\frac{g(t_n)}{f(t_n)}\to\beta$ for $n\to\infty$. Then

$$\frac{D_g\left(\left(\frac{t_n}{2},1-\frac{t_n}{2}\right),(1/2,1/2)\right)}{D_f\left(\left(\frac{t_n}{2},1-\frac{t_n}{2}\right),(1/2,1/2)\right)}\to\beta$$

and the first result follows.

The other three cases follows by interchanging f and g, and/or replacing f by f^* and g by g^*. We have used that

$$\lim_{t\to 0}\inf\frac{g^*(t)}{f^*(t)}=\lim_{t\to 0}\inf\frac{tg\left(t^{-1}\right)}{tf\left(t^{-1}\right)}=\lim_{t\to\infty}\inf\frac{g(t)}{f(t)}.$$

■

Proposition 10: Assume that f and g are C^2 and that $f''(1)>0$ and $g''(1)>0$. Assume that $\lim_{t\to 0}\inf\frac{g(t)}{f(t)}>0$, and that $\lim_{t\to\infty}\inf\frac{g(t)}{f(t)}>0$. Then there exists $\beta>0$ such that

$$D_g(P,Q)\geq\beta\cdot D_f(P,Q) \qquad (10)$$

for all distributions P,Q.

Proof: The inequality $\lim_{t\to 0}\inf\frac{g(t)}{f(t)}>0$ implies that there exist β_0,$t_0>0$ such that $g(t)\geq\beta_0 f(t)$ for $t<t_0$. The Inequality $\lim_{t\to\infty}\inf\frac{g(t)}{f(t)}>0$ implies that there exists $\beta_\infty>0$ and $t_\infty>0$ such that $g(t)\geq\beta_\infty f(t)$ for $t>t_\infty$. According to Taylor's formula we have

$$\begin{aligned}f(t)&=\frac{f''(\theta)}{2}(t-1)^2,\\ g(t)&=\frac{g''(\eta)}{2}(t-1)^2\end{aligned}$$

for some θ and η between 1 and t. Hence

$$\frac{g(t)}{f(t)} = \frac{f''(\theta)}{g''(\eta)} \to \frac{f''(1)}{g''(1)} \text{ for } t \to 1.$$

Therefore there there exists $\beta_1 > 0$ and an interval $]t_-, t_+[$ around 1 such that $\frac{g(t)}{f(t)} \geq \beta_1$ for $t \in]t_-, t_+[$. The function $t \to \frac{g(t)}{f(t)}$ is continuous on the compact set $[t_0, t_-] \cup [t_+, t_\infty]$ so it has a minimum $\tilde{\beta} > 0$ on this set. Inequality 10 holds for $\beta = \min\left\{\beta_0, \beta_1, \beta_\infty, \tilde{\beta}\right\}$. ■

IV. Examples

In this section we shall see a number of examples of how the method developed i this paper can be applied to determine the joint range for some pairs of f-divergences. Some of these results are known and others are new. We will not spell out all the details but shall restrict to the main flow of the argument that will lead to the joint range.

A. Power divergence of order 2 and 3

We have

$$f(t) = \phi_2(t),$$
$$g(t) = \phi_3(t).$$

In this case we have

$$D_f((p, 1-p), (q, 1-q)) = \frac{1}{2}\left(\frac{(p-q)^2}{q} + \frac{(p-q)^2}{1-q}\right),$$
$$D_g((p, 1-p), (q, 1-q)) = \frac{1}{6}\left(\left(\frac{p}{q}\right)^3 q + \left(\frac{1-p}{1-q}\right)^3 (1-q) - 1\right).$$

First we determine the image of the triangle. The derivatives are

$$\frac{\partial D_f}{\partial p} = \frac{2}{2} \cdot \frac{(p-q)}{(1-q)\, q} \ ,$$
$$\frac{\partial D_f}{\partial q} = \frac{1}{2} \cdot \frac{(2pq - q - p)(p-q)}{(1-q)^2 q^2} \ ,$$
$$\frac{\partial D_g}{\partial p} = \frac{-3}{6} \cdot \frac{(2pq - q - p)(p-q)}{(1-q)^2 q^2} \ ,$$
$$\frac{\partial D_g}{\partial q} = \frac{2}{6} \cdot \frac{\begin{pmatrix} pq + p^2 + q^2 - \\ 3pq^2 - 3p^2 q + 3p^2 q^2 \end{pmatrix}(p-q)}{(q-1)^3 q^3} \ .$$

The determinant of derivatives is

$$\begin{vmatrix} \frac{\partial D_f}{\partial p} & \frac{\partial D_g}{\partial p} \\ \frac{\partial D_f}{\partial q} & \frac{\partial D_g}{\partial q} \end{vmatrix} =$$
$$\frac{(p-q)^2}{12 q^4 (1-q)^4} \begin{vmatrix} 2 & 3p + 3q - 6pq \\ 2pq - q - p & \begin{pmatrix} 6pq^2 - 2p^2 - 2q^2 \\ -2pq + 6p^2 q - 6p^2 q^2 \end{pmatrix} \end{vmatrix}$$
$$= -\frac{1}{12}\left(\frac{p-q}{q(1-q)}\right)^4.$$

We see that the determinant of derivatives is different from zero for $p \neq q$ so the interior of Δ is mapped one-to-one to the image. Hence we just have to determine the image of points on the boundary of Δ (or near the boundary if undefined on the boundary).

For $P = (1, 0)$ and $Q = (1-q, q)$ we get

$$D_f(P, Q) = \frac{1}{2}\left(q + \frac{q^2}{1-q}\right) = \frac{1}{2}\left(\frac{1}{1-q} - 1\right),$$
$$D_g(P, Q) = \frac{1}{6}\left(\frac{1}{(1-q)^2} - 1\right) = \frac{1}{6}\frac{(2-q)\, q}{(1-q)^2}.$$

The first equation leads to

$$q = \left(1 - \frac{1}{2D_f + 1}\right)$$

and hence

$$D_g = \frac{2}{3} D_f (D_f + 1).$$

We have

$$\frac{f(t)}{g(t)} = \frac{\frac{t^2 - 2(t-1) - 1}{2}}{\frac{t^3 - 3(t-1) - 1}{6}} \to \infty \text{ for } t \to \infty.$$

All points $(0, s), s \in [0, \infty)$ are in the closure of the range of $(P, Q) \to (D_f, D_g)$. By combing these two results we see that the range consists of the point $(0, 0)$, all points on the curve $\left(x, \frac{2}{3}x(x+1)\right), x \in (0, \infty)$, and all point above this curve.

Similar results holds for any pair of power divergences, but for other pairs than (D_2, D_3) the computations become much more involved.

Note that the Rényi divergences are monotone functions of the power divergences so our results easily translate into the results on Rényi divergences. More details on Rényi divergences can be found in [13].

B. Total variation and χ^2-divergence

In this case we have

$$f(x) = |x - 1|,$$
$$g(x) = \frac{1}{2}(x-1)^2.$$

The function f is not differentiable but on the triangle Δ we have $p \leq q$ and

$$D_f(P, Q) = q\left|\frac{p}{q} - 1\right| + (1-q)\left|\frac{1-p}{1-q} - 1\right|$$
$$= 2(q-p).$$

Hence $D_f(P, Q)$ is C^∞ on Δ although f is not differentiable. We get

$$\frac{\partial D_f}{\partial p} = -2 \ ,$$
$$\frac{\partial D_f}{\partial q} = 2 \ ,$$
$$\frac{\partial D_g}{\partial p} = \frac{(p-q)}{(1-q)\, q} \ ,$$
$$\frac{\partial D_g}{\partial q} = \frac{(2pq - q - p)(p-q)}{2(1-q)^2 q^2} \ .$$

Hence

$$\left|\begin{array}{cc} \frac{\partial D_f}{\partial p} & \frac{\partial D_g}{\partial p} \\ \frac{\partial D_f}{\partial q} & \frac{\partial D_g}{\partial q} \end{array}\right| = \left|\begin{array}{cc} -2 & 2 \\ \frac{(p-q)}{(1-q)q} & \frac{(2pq-q-p)(p-q)}{2(1-q)^2 q^2} \end{array}\right|$$
$$= -2\frac{(q-p)^2 (q-1/2)}{(1-q)^2 q^2}.$$

The mapping Δ to the range of (D_f, D_g) is singular for $q = 1/2$. The line $p \to (p, 1/2)$ is mapped into the curve

$$p \to (D_f(P,Q), D_g(P,Q))$$
$$= \left(2\left(p-\frac{1}{2}\right), 2(p-1/2)^2\right).$$

If the total variation is denoted V this curve satisfies $\chi^2 = \frac{1}{2}V^2$ and points satisfying $\chi^2 \geq \frac{1}{2}V^2$ are 2-achievable. The inequality $\chi^2 \geq \frac{1}{2}V^2$ has been proved previously by a different method [14].

C. Total variation and LeCam divergence

On the triangle Δ we have

$$D_f(P,Q) = 2(q-p),$$
$$D_g(P,Q) = \frac{1}{4}\left(\frac{(p-q)^2}{p+q} + \frac{(p-q)^2}{2-p-q}\right).$$

The derivatives of the LeCam divergence is

$$\frac{\partial}{\partial p} D_g(P,Q) = \frac{(p-q)\left(p+3q-2pq-2q^2\right)}{(p+q)^2 (2-p-q)^2},$$
$$\frac{\partial}{\partial q} D_g(P,Q) = \frac{\left(2pq-q-3p+2p^2\right)(p-q)}{(p+q)^2 (p+q-2)^2}.$$

Hence

$$\left|\begin{array}{cc} \frac{\partial D_f}{\partial p} & \frac{\partial D_g}{\partial p} \\ \frac{\partial D_f}{\partial q} & \frac{\partial D_g}{\partial q} \end{array}\right|$$
$$= \left|\begin{array}{cc} -2 & 2 \\ \frac{(p-q)\left(p+3q-2pq-2q^2\right)}{(p+q)^2(2-p-q)^2} & \frac{\left(2pq-q-3p+2p^2\right)(p-q)}{(p+q)^2(p+q-2)^2} \end{array}\right|$$
$$= \frac{4(1-p-q)(q-p)^2}{(p+q)^2 (p+q-2)^2}.$$

The mapping is singular for $q = 1-p$. We get the curve

$$p \to \left(2(p-(1-p)), \frac{(p-(1-p))^2}{p+(1-p)} + \frac{(p-(1-p))^2}{2-p-(1-p)}\right)$$
$$= \left(4\left(p-\frac{1}{2}\right), 2\left(p-\frac{1}{2}\right)^2\right).$$

If total variation is denoted V then the curve is $D_g = \frac{1}{8}V^2$ and any point above this curve is achievable.

D. Information divergence and reversed information divergence

In this case we have

$$f(t) = t\ln t,$$
$$g(t) = -\ln t.$$

We see that $g(0) = \infty$ and that $\frac{g(t)}{f(t)} \to \infty$ for $t \to 0$. Lemma 9 implies that the supremum of

$$D_g(P,Q) - \gamma D_f(P,Q) = D(Q\|P) - \gamma D(P\|Q)$$

over all distributions P, Q is ∞ for any $\gamma < \infty$. Similarly the supremum of

$$D(P\|Q) - \gamma D(Q\|P)$$

over all distributions P, Q is ∞ for any $\gamma < \infty$. Since $(0,0)$ is in the range and the range is convex, the range consist of all interior points of the first quadrant and the point $(0,0)$.

V. Acknowledgement

The authors thank Job Briët and Tim van Erven for comments to a draft of this paper.

This work was supported by the European Network of Excellence and the GAČR grants 102/07/1131 and 202/10/0618.

References

[1] I. Csiszár, "Eine informationstheoretische Ungleichung und ihre anwendung auf den Beweis der ergodizität von Markoffschen Ketten," *Publ. Math. Inst. Hungar. Acad.*, vol. 8, pp. 95–108, 1963.

[2] T. Morimoto, "Markov processes and the h-theorem," *J. Phys. Soc. Jap.*, vol. 12, pp. 328–331, 1963.

[3] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Roy. Statist. Soc. Ser B*, vol. 28, pp. 131–142, 1966.

[4] F. Liese and I. Vajda, *Convex Statistical Distances*. Leipzig: Teubner, 1987.

[5] F. Liese and I. Vajda, "On divergence and informations in statistics and information theory," *IEEE Tranns. Inform. Theory*, vol. 52, pp. 4394 – 4412, Oct. 2006.

[6] T. R. C. Read and N. Cressie, *Goodness of Fit Statistics for Discrete Multivariate Data*. Berlin: Springer, 1988.

[7] A. R. Barron, L. Györfi, and E. C. van der Meulen, "Distribution estimates consistent in total variation and in two types of information divergence," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1437–1454, Sept. 1992.

[8] A. Fedotov, P. Harremoës, and F. Topsøe, "Refinements of Pinsker's inequality," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1491–1498, June 2003.

[9] J. Briët and P. Harremoës, "Properties of classical and quantum Jensen-Shannon divergence," *Physical review A*, vol. 79, p. 052311 (11 pages), May 2009.

[10] F. Topsøe, "Bounds for entropy and divergence of distributions over a two-element set," *J. Ineq. Pure Appl. Math.*, 2001. [ONLINE] http://jipam.vu.edu.au/accepted_papers/ 04400.html.

[11] R. R. Phelps, *Lectures on Choquet's Theorem*. No. 1757 in Lecture Notes in Mathematics, Springer, second edition ed., 2001.

[12] V. Boltyanski and H. Martini, "Carathodory's theorem and H-convexity," *Journal of Combinatorial Theory, Series A*, vol. 93, no. 2, pp. 292 – 309, 2001.

[13] T. van Erven and P. Harremoës, "Rényi divergence and majorization," in *Proceedings ISIT 2010*, pp. 1335–1339, IEEE, June 2010.

[14] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, pp. 419–435, 2002.